Exon-Intron Boundary Detection Made Easy by Physicochemical Properties of DNA

B. Jayaram

bjayaram@chemistry.iitd.ac.in

Indian Institute of Technology Delhi https://orcid.org/0000-0002-5495-2213

Dinesh Sharma

Indian Institute of Technology Delhi https://orcid.org/0009-0002-6997-8683

Danish Aslam

Indian Institute of Technology https://orcid.org/0000-0001-5372-449X

Kopal Sharma

Indian Institute of Technology Delhi

Aditya Mittal

Indian Institute of Technology Delhi https://orcid.org/0000-0002-4030-0951

Article

Keywords:

Posted Date: May 27th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4359229/v1

Additional Declarations: There is **NO** Competing Interest.

Abstract

Genome architecture in eukaryotes exhibits a high degree of complexity. Amidst the numerous intricacies, the existence of genes as non-continuous stretches composed of exons and introns has garnered significant attention and curiosity among researchers. Accurate identification of exon-intron boundary junctions is crucial to decipher the molecular biology governing gene expression of regular and aberrant splicing. The currently employed frameworks for genomic signals, which aim to identify exons and introns within a genomic segment, need to be revised primarily due to the lack of a robust consensus sequence and the limitations posed by the training on available experimental data sets. To tackle these challenges and capitalize on the understanding that deoxyribonucleic acid (DNA) exhibits function-dependent local structural and energetic variations, we present ChemEXIN, an innovative method for predicting exon-intron boundaries. The method utilizes a deep-learning (DL) model alongside tri- and tetra-nucleotide-based structural and energy parameters. ChemEXIN surpasses current methods in accuracy and reliability. Our work represents a significant advancement in exon-intron boundary annotations, with potential implications for understanding gene expression, regulation, and biomedical research.

Introduction

In the heterogenous world of genomics, eukaryotes stand apart from prokaryotes with a fascinating twist – their genetic blueprints exhibit remarkable complexity[1]. Amongst various captivating elements in eukaryotic DNA, the intriguing exon-intron boundary regions have ignited a blazing spark of interest among researchers.

A gene in eukaryotes is a discontinuous structure composed of a protein-coding region (exon) and a non-coding stretch (intron)[2]. During the process of gene expression, the introns are excised from a pre-messenger ribonucleic acid (pre-mRNA) after transcription, and the exons are joined together through splicing in various combinations to form mature mRNA products[3]. These exon-intron boundary sites are vital for determining the encoded amino acid sequence and for regulating splicing events. These boundaries hold significant medical importance, as many human genetic disorders and diseases result from irregular pre-mRNA splicing[4]. Thus, the demarcation of accurate exon-intron architecture is crucial in eukaryotic genome annotation.

In pursuit of annotating these sites, several attempts have been made in genomics. In the early stages of exploration, researchers relied upon the consensus sequence-based approach[5,6]. Scrutinizing the sequences, character by character, and complementing the findings with the experimental data provided with the initial patterns for their identification. These signals, generally known as splice site (SS) motifs, occur in nucleotide pairs with GT and AG at the 5’ and 3’ ends of the intron, respectively[7,8]. However, at later stages, the emergence of cryptic SSs within all the genes of a particular eukaryotic species and other organisms yielded several diverse consensus stretches[9,10]. The situation is even more complex due to the prevalence of alternative splicing (AS) in eukaryotes. An individual gene can give rise to multiple mRNA isoforms through AS by selectively including or excluding different exons, creating an array of potential protein products[11]. This remarkable phenomenon adds another layer of complexity to the identification of exon-intron boundaries, as the traditional linear gene model no longer suffices.

Researchers have recognized the need for a more comprehensive and reliable approach. Various computational approaches have long been used to annotate these evasive boundaries. Specific tools leverage scoring matrices holding valuable sequence pattern information from experimentally verified SSs by identifying conserved nucleotide positions and their frequencies[12,13,14,15]. Approaches like Genscan[16] and GenomeScan[17] incorporate additional information from known protein sequences to enhance their predictive power. Advanced algorithms, such as GeneWise[18], Augustus[19], Fgenesh[20], GeneParser[21], and geneid[22] are built using dynamic programming models employing a data-driven approach to learn the sequence patterns associated with various genomic elements. Spliceator[23], a recent innovative approach to splice site prediction, harnesses the power of convolutional neural network (CNN) for its predictive capabilities. The key strength of

Spliceator lies in its training process, which uses validated data from a diverse set of over 100 organisms. While these methods demonstrate substantial predictive capabilities, their effectiveness relies heavily on the availability of extensive sequence data, resulting in variable performance from species to species.

In addition to the aforementioned methods, ribonucleic acid (RNA) based tools offer reliable predictions for organisms with or without a reference genome. Unfortunately, these tools, too, fall short when it comes to annotating splice junctions in DNA sequences[24,25,26,27,28]. Recent exploration of chromatin organization and nucleosome positioning approach presents a fresh perspective[29]. It has yet to achieve the desired level of sensitivity and specificity. Although valuable insights have been gathered from these studies over the years, it remains clear that novel ideas and newer models are essential for accurately identifying exon-intron boundaries in genome sequences.

It is widely recognized that DNA within our body exhibits sequence and more importantly function-specific local structural and energetic variations[30,31,32,33,34,35,36,37,38]. These arrangements are necessary to facilitate several biological processes, such as protein interactions, gene expressions, etc.[39]. Investigations on nucleic acid structures have yielded fresh insights into genome architecture, providing researchers with a new perspective on annotation. Consistent findings from studies demonstrate that similar DNA sequences often share similar biophysical properties. Interestingly, however, it is not always the case that alternative sequences can produce DNA molecules that possess similar structures and energy properties[40,41]. This intriguing phenomenon highlights the complex relationship between DNA sequences and their resulting physicochemical properties.

Working with these physicochemical properties, our past research has highlighted the significance of biophysical profiling in the characterization and annotation of DNA elements[8,42,43,44,45,46,47,48,49,50]. These findings reveal that the physicochemical signatures of genomic elements are unique and conserved despite sequence variations at these sites. In line with this trend, exon-intron boundaries also display distinctive structural and energy profiles, distinguishing them from other genomic regions[48]. Advancing our exploration into eukaryotic genome annotation, we present a novel approach, ChemEXIN, which utilizes structural and energy characteristics of DNA to identify exon-intron boundaries. This method capitalizes on Molecular dynamic (MD) based biophysical features encompassing the Backbone, Base pair (BP) axis organization, Inter-BP organization, Intra-BP organization, and energetics of DNA to discern the precise exon-intron boundary junctions.

ChemEXIN has undergone dedicated training and development on all exon-intron boundary junctions from the protein-coding genes in *Homo sapiens* (*H. sapiens*), *Mus musculus* (*M. musculus*), and *Caenorhabditis elegans* (*C. elegans*). It is openly accessible on GitHub and has been extensively optimized using comprehensive datasets involving rigorous comparisons vis a vis various classification models. Further, comparing it against widely adopted sequence-based methods using exhaustive datasets complements its versatility. These findings underscore the robustness and broad applicability of ChemEXIN in accurately predicting exon-intron boundaries across both protein-coding and non-coding genes. This comprehensive performance not only validates the effectiveness of our approach but also positions it as a significant contributor to the progression of eukaryotic gene annotation methodologies.

Results

1) Physicochemical profiles at the exon-intron boundaries

Figure 1A-1E, **Figure S1, and Figure S2 (Supplementary File 4)** depict the numerical profiles of all 28 parameters, including the nine Backbone angle parameters, six Inter-BP parameters, six Intra-BP parameters, four BP-axis parameters, and three energy parameters. These profiles were generated for coding sequences (CDS), exon-start (ES), and exon-end (EE) sequences in *H. sapiens*. The results of this study indicate that the physicochemical profiles linked to each parameter in the

five main categories for both ES and EE sequences display unique patterns, which differ significantly from those observed in CDSs. The results demonstrate that while the structural and energy properties of the CDS remain

relatively constant throughout the sequences, a distinct shift occurs in the biophysical profiles at the exon-intron boundary within sequences harboring them. The structural trends observed at the exon-intron sites in Fig. 1A-1D, for each parameter, emphasize the presence of a transient thermodynamically unstable boundary. This boundary is crucial for facilitating classical splicing events, with exons demonstrating higher stability than neighboring intronic sites[51,52]. Additionally, the energy plots in Fig. 1E, at the exon-intron junctions within these DNA sequences, support the classical hypothesis that boundary elements play a crucial role in secondary structure formation in RNA, thereby facilitating splicing. The Hydrogen bond energy exhibited a rapid rise followed by a drop, implying an initial instability at the boundary position that gradually balances out as the junction site progresses. In contrast, the Stacking energy reached its maximum value at the border junction, leading to an increased flexibility in the DNA by reducing its stiffness. The observed decreased Solvation energy could indicate the transiently formed stable structure at the interfaces between exons and introns.

Moving further with the idea that the combined effect of smaller features brings about a concerted change, we combined the individual structural parameters belonging to the respective major categories to provide us with the actual Backbone, Inter-BP, Intra-BP, and BP-axis profiles. The synergistic visualization of these categories and the three energy parameters at the exon junctions are available in Fig. 1F. These results provide us with the evident change at the boundaries for the seven structural and energy parameters. The trend, initially widespread over a region of 50–100 length within the individual parameters, is now contained uniformly within a region of ~ 50 for all the categories. The shaded region within the combined plots shows the site undergoing major structural and energy changes. Together, these individual and combined profiles offer valuable insights into the potential utility of the combined parameters for effective exon-intron boundary identification within any given gene sequence.

2) Correlation analyses and feature importance

A correlation analysis was conducted to examine the interrelationships among the seven final parameters within the 50-nucleotide regions of both the ES and EE profiles in humans. The primary objective of these analyses was to assess the degree of correlation between parameters and identify any redundancy that may exist. Figure 2 shows the correlation results. Different pairs of parameters exhibited varying degrees of correlation, ranging from moderate to high. Some parameters were dependent on each other, while others showed no correlation. Furthermore, an examination of feature importance through principal component analysis (PCA) was performed to retain the significant features without compromising on information for the downstream analysis. As summarized in Fig. 2 and **Methodology S1 (Supplementary File 2)**, these results emphasize the significance of using all seven parameters. The methodology, as outlined in Fig. 3, was thus followed, leading to the development of the novel physicochemical property-based exon-intron boundary prediction method, ChemEXIN.

Figure 2. **(A)** Correlation matrices depicting the relationships among the seven final parameters at exon-start and exon-end. **(B)** Feature importance analysis conducted through PCA, revealing significant contributions of all seven parameters.

3) Performance evaluation

To arrive at an optimal exon-intron boundary prediction method, various Machine Learning (ML)/DL models were deployed during the initial development phases of ChemEXIN. The performance of these models underwent comparison using both the training-testing split dataset and the evaluation dataset in humans. Model assessment and comparison were conducted using five key criteria: sensitivity, specificity, F1-score, precision, and accuracy. The conclusive results of the training-testing are presented in **Table S1 (Supplementary File 3)**. These results indicate that all the models exhibit the capability to predict exon-intron boundary sites, with accuracy levels and F1-scores spanning from 54% when utilizing basic models to an improved performance of ~ 80% when employing DL model on the test set. On the same dataset, it is worth

noticing that parameters such as specificity, which examines the model's ability to accurately detect true negatives, and sensitivity (true positive rate or recall), which evaluates the system's proficiency in predicting true positives within each category or class, demonstrated notably strong performance values.

The findings from the *H. Sapiens* evaluation set comprising 60,000 held-out sequences further validate the efficacy of utilizing biophysical parameters for accurately identifying boundary regions **(Table S1, Supplementary File 3)**. Further, the area under the receiver operating characteristic curve (AUROC) scores on the evaluation dataset presented in Fig. 4 show that the three-dimensional CNN (3D-CNN) and Support Vector Machine (SVM) classifiers surpass other models in predictive performance across all three classes. Notably, the 3D-CNN exhibits higher area under the curve (AUC) values across all classes, signifying its efficacy in distinguishing between diverse classes. Following the comparison results of the above models and the three-dimensional nature of our datasets, we decided to proceed with the 3D-CNN[53] trained model for subsequent analysis and its independent implementation in all the three organisms under study. The architecture of the 3D-CNN model employed here is detailed in Fig. 5 **and Methodology S2 (Supplementary File 2).**

4) Comparison with the state-of-the-art tools

Five widely used gene structure organization prediction tools —Spliceator, Fgenesh, geneid, Genscan, and Augustus, were benchmarked against each of our three organism-specific trained models. The results are presented in Fig. 6 **and Tables S3, S4, and S5 (Supplementary File 3)**, with details on the outputs available in **Methodology S3 (Supplementary File 2).** To ensure an unbiased comparison of our approach, we used three benchmarking datasets, each comprising 2,000 randomly selected sequences from the respective organism. The majority of these tools are available as web servers, which tend to crash on large input sequences and/or require input sequences in batches. Henceforth, this reasonable-sized comparison data ensured the efficient working of all the tools.

Spliceator, available as a web server[23], employs CNN in conjunction with a user-defined reliability parameter and a sequence search window to predict the gene organization within input sequences. Instead of treating individual input sequences separately, it processes them as a unified input string with a maximum length of approximately 200,500 bases. To adhere to this constraint, we divided the input sequences for ES and EE for organisms under study into two batches. We employed a default reliability parameter score of 98% and a model tailored to a 400-length search window (as our individual input sequences are 401 nucleotides long) to predict donor and acceptor sites. The output files obtained for each batch were combined into their respective categories and processed to provide a final confusion matrix. From the results, it is evident that Spliceator results are less than satisfactory for all three organisms. The observed high level of misclassification is primarily attributed to Spliceator's consensus-based approach to identifying donor and acceptor sites, resulting in an over-representation of these sites in the predictions. This over-representation tends to increase with a decrease in the reliability parameter score due to non-specific pattern matching. Moreover, there is no noticeable improvement at a 100% reliability parameter score, suggesting its high sequence specificity.

Fgenesh is available both as a web server and as a local downloadable version. Due to the requirement of several genomic feature files in processing the downloadable version, we tested our sequence with the online web server[20]. The method accepts a single file as input, with each sequence represented in FASTA format. In addition to a Hidden Markov model (HMM) based gene prediction model trained over several eukaryotic species, though not within the scope of our research, it provides numerous user-specific advanced options. Operating it with organism-specific default parameters, the method generated output files, which were then processed into a confusion matrix. Similar to Spliceator, Fgenesh yielded comparable results for *C. elegans* and *M. musculus*. However, for humans, the precision and accuracy notably improved, reaching close to 40%. Although there is a potential for improved outcomes by utilizing targeted training with specific feature files in the downloadable version (Fgenesh++), we opted not to pursue these advanced options.

geneid, another tool in our evaluation, employs position weight arrays, scoring, and Markov models to identify gene features in DNA sequences. Although available as a web server and a GitHub repository[22], we faced challenges with the online

version, prompting us to resort to the local version downloaded from GitHub[22]. Despite processing input sequences in a manner similar to Fgenesh, the processed results more closely resemble those of Spliceator, exhibiting a high misclassification rate ranging from 80–90% for the organisms under consideration. The misclassification observed can be attributed to the overrepresentation of the ES and EE sites. Regardless of being trained on multiple species from all four eukaryotic kingdoms, geneid did not yield satisfactory results in our study.

Continuing our benchmarking efforts, we evaluated Genscan, a widely used tool for identifying exon-intron structures in genomic sequences. Genscan[16] employs general probabilistic models to annotate gene features within input sequences. While Genscan can process nearly one million bases, our dataset, comprising approximately 0.8 million bases, posed a challenge to its processing capabilities. Hence, following a similar strategy employed with Spliceator, we partitioned the input sequences into two batches for both the ES and EE. Regrettably, akin to the outcomes observed with the other tools, the results were not encouraging.

Transitioning to our last tool, Augustus, our objective was to evaluate its proficiency in predicting gene structures. Beyond being accessible as a straightforward pre-trained web server and a GitHub repository, Augustus offers an improved web server option. This server allows training sequences not listed in their database using annotation files containing information for complementary DNA (cDNA) sequences and/or hints for donor and acceptor sites (hint files). We utilized the pre-trained web server[19] and prepared a basic hint file (GTF) for input sequences (FASTA), adhering to the required format. Using a generalized HMM with an additional probabilistic model for gene structure prediction, Augustus also provides information on alternate SSs. Augustus exhibited relatively favorable performance compared to other tools, achieving a specificity of approximately 85%, notably attributed to the utilization of a hint file. In the case of humans, the misclassification rate decreased significantly to approximately 45%. However, while a similar trend was observed for other organisms, the outcomes were less favorable.

In a similar manner to the above-reported comparisons, we examined how well our models performed by looking at their predictions on the benchmarking datasets. This evaluation was essential for understanding how accurately our approach could predict exon-intron boundaries. Our analysis unequivocally demonstrates that our approach clearly outperforms other tools (Fig. 6) across all major evaluation criteria in all three organisms. The results indicate a notably low misclassification rate, ranging approximately from 0.075 to 0.20, and high precision, ranging from approximately 0.796 to 0.92. These findings indicate the reliability and accuracy of the predictions obtained through our technique. This exhaustive comparison underscores the presence of substantial sequence alternatives. However, despite these variations, the biophysical profiles at the junction sites remain largely conserved. This conservation suggests the potential utility of these profiles in facilitating precise recognition and prediction by our physicochemical property-driven 3D-CNN models.

Expanding the scope of our comparison, we further assessed the performance of the reported method alongside two top-performing tools identified in the previous benchmarking step, namely Fgenesh and Augustus. This extended comparison focused on predicting exon-intron junctions in non-protein coding genes, including long-non-coding RNA (lncRNA) genes; transfer RNA (tRNA) genes; and ribosomal RNA (rRNA) genes in humans.

Despite its widespread usage, Fgenesh failed to generate results in our comparative assessment. Unlike Augustus, while our method has not undergone specific training on the exon-intron characteristics of these genes, the results documented in **Table 1 and** Fig. 7 underscore a notable performance of our framework against Augustus, indicating its adaptability and efficacy even in contexts beyond the specialized training domain. This underscores the robustness and versatility of our approach, particularly in addressing gene prediction tasks across varied genomic contexts.

Leveraging biophysical parameters and the DL method, our approach exhibited superior performance compared to existing gene annotation tools across the three organisms. Moving forward, we developed ChemEXIN, a consolidated prediction framework combining the three organism-specific models. This approach holds significant potential for enhancing the efficiency of exon-intron boundary annotation.

5) Exon-intron boundary prediction through ChemEXIN

ChemEXIN, available as an open-source tool, can be downloaded and used within a conda environment, offering an accessible platform for researchers. After the initial setup of the virtual environment through

cloning, users can activate and run ChemEXIN using a Python 3 interpreter via a command prompt. This process involves providing essential inputs: a file containing the gene sequence of interest, the associated

organism, and a threshold value that defines the probability at which prediction windows are refined. Upon receiving these inputs, ChemEXIN performs its analysis and delivers the prediction results in a comma-delimited file. For detailed instructions on setting up and using ChemEXIN, researchers can refer to the user manual **(Supplementary File 5)**.

To assess the performance of ChemEXIN, we tested it on random gene sequences of varying lengths from the studied organisms, using a default probability score of 0.75. The specific outcomes of this analysis are cataloged in Table 2. Additionally, to assess ChemEXIN's compatibility across different computing environments, we executed predictions on the same gene set but on systems with various configurations. The results of this compatibility assessment are detailed in **Table S6 (Supplementary File 3)**. Collectively, these results demonstrate that ChemEXIN is highly efficient in processing sequences of diverse lengths, a feat it accomplishes using minimal computational resources and without depending on the Operating system (OS). A detailed examination of our prediction outcomes, particularly with human and mouse gene sequences, reveals that a significant number of boundary sites are predicted with remarkable accuracy. Even in many instances where predictions deviate, they do so by a margin of only five to ten nucleotides from the established boundary windows. Notably, Despite the promising results from training-testing and benchmarking analyses, the predictions made by the *C. elegans* model were not reliable and, hence, not reported. This irregularity could be attributed to inadequate training from imbalanced positive and negative datasets. Enhancing the accuracy for *C. elegans* thus requires refining the filters and extending the training process. These improvements will yield better results and broaden the applicability of ChemEXIN across various eukaryotes, paving the way for future advancements.

Table 2
Performance evaluation of ChemEXIN on random genes from H. sapiens and M. musculus.

Organism	Gene	Length (nt)	Predicted Sites	Average time (sec)
H. sapiens	*DMD*	2,220,382	47	221.77
	BDNF	188,307	28	24.84
	NEU1	10,881	8	8.24
M. musculus	*RP1*	409,685	26	41.30
	CDK6	189,524	9	22.46
	SCAF8	83,888	15	12.78

Discussion

Our study extensively examined the structural and energy profiles at exon-intron boundaries in humans through a comprehensive analysis of 28 biophysical parameters. This investigation revealed distinct physicochemical patterns at the exon-intron boundaries, which are markedly different from relatively stable profiles observed within CDSs. The patterns at the exon-intron boundaries are crucial in facilitating classical splicing events and can be utilized for their recognition. Building on this foundation, our correlation analyses and feature importance assessments highlighted the synergistic effect of the physicochemical parameters in defining the exon-intron boundaries. By aggregating individual parameters into broader categories, we successfully encapsulated the changes across these boundaries within a compact region of

approximately 50 nucleotides. This methodology refined our understanding of identifying these boundaries and underscored the potential of physicochemical properties in enhancing the annotation.

The development and subsequent evaluation of our prediction methodology commenced by integrating two additional eukaryotic model organisms. By leveraging a blend of biophysical parameters and DL, our models notably outperformed existing gene organization prediction tools across various evaluation metrics in all the three organisms studied. This achievement illustrated the effectiveness of combining structural and energetic profiles with sophisticated computational methods for boundary prediction. Further comparison with state-of-the-art tools emphasized the limitations of current methods in accurately identifying exon-intron junctions, particularly in non-protein coding genes. Despite these challenges, the robust performance of our methodology highlights its adaptability and potential for broader applications in gene prediction tasks, even in contexts where traditional tools fail.

Finally, integrating the prediction models with refinement filters, we made ChemEXIN available as an open-source tool, embodying a user-friendly interface that makes it a valuable resource for the research community. Its ability to efficiently process diverse gene sequences, its minimal computational demands, and compatibility across different systems, significantly enhance its utility in genomic studies. This seamless integration of comprehensive analysis, innovative methodology, and accessible technology is pivotal in understanding and predicting exon-intron boundaries, setting a new benchmark for future genomic research.

Undeterred by its utility in predicting exon-intron boundaries, ChemEXIN exhibits limitations that may affect its performance and applicability in specific contexts. The first intrinsic limitation of ChemEXIN arises from the identical biophysical characteristics at the ES and EE sites. This issue, highlighted by the structure and energy graphs in Fig. 1 and further supported by AUROC curves in Fig. 4, makes it difficult to differentiate them. The final prediction pipeline is thus tailored to identify boundary occurrences collectively rather than distinguishing between individual ES and EE sites. Further, due to the absence of biophysical characters for entire genomic regions, ChemEXIN demonstrates efficacy in predicting exon-intron boundaries within gene sequences. Its performance diminishes notably when analyzing full-length genome sequences. Additionally, while ChemEXIN effectively considers standard splicing events, it does not account for AS variants, limiting its ability to characterize alternate SSs. A potential solution lies in exploring boundary-like patterns occurring at non-traditional sites within the unfiltered raw output files. However, this is outside the scope of the current investigation. Finally, despite delivering promising results in training-testing and benchmarking steps, ChemEXIN's performance on *C. elegans* could have been more optimal, highlighting the need for further investigation and refinement.

In conclusion, ChemEXIN, through biophysical parameters and 3D-CNN models, outperforms existing gene organization prediction tools, demonstrating its adaptability and potential for broader applications in genomic research. The identification of distinct physicochemical patterns at exon-intron boundaries highlights the importance of considering boundary-specific characteristics for accurate prediction. ChemEXIN's integration with refinement filters enhances its usability, offering a user-friendly platform for researchers. Regardless of certain limitations, such as difficulty in distinguishing between ES and EE sites and reduced performance with full-length genome sequences, future efforts could focus on refining the tool to ensure robust performance across diverse organisms. Overall, ChemEXIN represents a significant advancement in genomic research, with implications for enhancing our understanding of gene architecture and facilitating precise exon-intron boundary annotations.

Methods

1) Exon-intron sequence datasets

From the human genome feature files downloaded from the GENCODE[54] database, we identified and filtered out ES and EE positions from all protein-coding genes (a total of 328,368). Using the human reference genome, we generated two positive sequence datasets around these positions for both ES and EE.

The Dataset I consist of 401 nucleotides long 328,368 sequences. These sequences were generated through an extraction of 200 nucleotides located both upstream and downstream of the EE, positioned at zero. Similarly, Dataset II was created by spanning 200 nucleotides upstream and downstream of the ES. A negative control dataset consisting of 30,140 sequences, each extending 401 nucleotides, was similarly created using the CDSs. These sequences were extracted from the middle of exons with a length greater than 1,000 nucleotides.

2) Characterization parameters

For a comprehensive structural depiction of DNA, we have considered various aspects of its organization, including the Backbone arrangement defined by alpha, beta, gamma, delta, epsilon, zeta, chi, phase, and amplitude; Inter-BP arrangements through shift, slide, rise, tilt, roll, and twist; Intra-BP arrangements encompassing shear, stretch, stagger, buckle, propel, and opening; and the BP-axis, which takes into account X-displacement, Y-displacement, inclination, and tip.

In contrast to our previous studies, which relied on X-ray-derived dinucleotide data[8,42,43], our current research adopts a more comprehensive approach. We incorporate neighboring effects by analyzing the structural attributes of all distinctive tri-nucleotides to obtain parameter values for the Backbone, Intra-BP, and BP-axis and unique tetra-nucleotide steps for the Inter-BP arrangement parameters. The Nucleotide Database (NDB) lacks B-DNA structures encompassing all possible tri- and tetra-nucleotide steps. Therefore, akin to our recently published study[48], we rely on atomistic MD simulations as the sole viable approach to obtain reliable and transferrable parameters for all the unique nucleotide steps. To obtain these structural parameters, we synthetically designed 13 oligomers and followed the exact methodology outlined in our previous work[48]. For the energy parameters, we relied upon our in-house lab software to calculate the values of Hydrogen bond energy, Stacking energy, and Solvation energy over all instances of tri-nucleotide steps[50].

After computing all structural and energy parameters for each oligomer, we assessed the tri- and tetra-nucleotide steps in the 5' to 3' direction corresponding to each property. By averaging these occurrences, we generated comprehensive parameter value tables (**Supplementary File 1**).

3) Exon-intron boundary junction profiling and visualization

Using the tri- and tetra-nucleotide parameter value tables, every sequence within each dataset (328,368 ES/EE sequences and 30,140 CDS sequences) was converted to 28 numerical profiles. To avoid noise, these numerical profiles were subjected to a sliding window of 25 base pairs. Within this window, the values were averaged, resulting in a single value for each position. The resulting 374 and 373 long numerical profiles corresponding to the tri- and tetra-nucleotide parameters represented the parameter trend over the sequence. Thereafter, a min-max normalization was applied over these profiles, ensuring all values fell within a standardized range of zero to one[48].

A visual representation of these profiles was achieved by creating two categories of plots for both the ES and EE parameters. In the first category, discrete properties belonging to a main structural class were plotted directly on a single graph and compared with the CDS profiles. In the second category, numerical profiles of the structural parameters within a specific class were combined to generate a single curve, highlighting the synergistic effects of parameters within that group. This approach allowed us to observe the overall trends in Backbone, BP-axis, Intra-BP, and Inter-BP organizations. The three energy parameters represent different aspects, so they were kept separate.

4) Formulation of training datasets

For both the ES and EE sequences, the combined seven numerical profiles were processed to extract a segment of length 50, ranging from position 158 to 207. These segments, during visualization, displayed a distinctive pattern to that of the CDS sequences and acted as target classes for our prediction models.

To incorporate contrasting features, the seven parameter profiles from the CDS sequences were generated. However, this time, we employed a slightly different approach to capture the sequence characteristics and eliminated bias arising from a

lower count of CDS sequences (30,140). We extracted seven non-overlapping numerical fragments, each 50 in length. This extraction followed an organized, non-redundant approach, starting from position one and advancing in increments of 50 nucleotides (e.g., from position one to position 50, position 51 to position 100, and so forth, ultimately resulting in the final fragment spanning from position 300 to position 350). Consequently, this method yielded a negative training dataset comprising 210,980 CDS sequences corresponding to the seven final parameters.

5) Training Pipeline

The entire approach employed for ChemEXIN is outlined as a flowchart in Fig. 3. Before advancing to the training phase, our analysis commenced with investigating the correlations among the final parameters. This preliminary step aimed to elucidate the relationships and potential interdependencies between the parameters, providing valuable insights into their collective behavior. Correlation analyses conducted on the 50-length segment for both ES and EE datasets yielded diverse levels of correlation among different pairs of parameters. Nonetheless, these correlations were not much pronounced, except for a few cases observed in both datasets. A feature importance analysis was performed to strengthen the conclusions further. Consequently, for the scope of this study, all seven individual parameters were considered, and the positive (ES/EE) and negative (CDS) datasets were integrated into a single training sequence file. To get a vigorous prediction pipeline, instead of averaging the 50 values extracted from the numerical profiles of each parameter, all 50 values corresponding to each of the seven primary categories were treated as distinct features. This method retained the full spectrum of information within each category and thus provided us with 350 derived features (50 numerical values corresponding to each category) for each sequence. Advancing towards the training process, the integrated dataset comprising ~ 850,000 sequences was categorized into three classes: 0 for CDS, 1 for ES, and 2 for EE. These sequences were then separated into smaller datasets for extensive training-testing and evaluation. 60,000 sequences having an equal proportion of CDS, ES, and EE were chosen randomly from their respective classes and constituted the blind evaluation dataset. The remaining sequences, after randomization, which ensured unbiasedness, were subjected to a classical 80 – 20 split to create training-testing datasets. Various ML/DL methods were deployed over these human datasets, and the results were compared. By employing multiple models rather than relying on a single one, we strengthened the idea that the physicochemical profiles observed at the exon-intron boundaries play a crucial role in predictions.

Starting here, we employed a parallel methodology for two additional eukaryotes, namely *M. musculus* and *C. elegans*. This approach involved a similar exon-intron sequence extraction and training-testing split alongside an independent extraction of a benchmarking set. This set comprised 2,000 sequences each for ES and EE for each organism **(Supplementary File 1)**. Skipping the organism-specific model evaluation, the best-performing model in humans was deployed for these organisms. Further, to maintain linearity across organisms arising from the evaluation dataset corresponding to *H. sapiens*, 2,000 random ES and EE sequences from this set were selected as a separate benchmarking set for humans.

6) Evaluation and comparison with the state-of-the-art

To assess our trained method, which includes models from *H. sapiens, M. musculus*, and *C. elegans*, we benchmarked it against five widely used gene annotation tools. This state-of-the-art comparison was conducted using the organism-specific benchmarking datasets.

Additionally, an advanced comparison between the two top-performing tools and the model was conducted using non-protein coding gene datasets. These datasets encompass sequences devoid of prior training, thus offering a rigorous evaluation of the efficacy and adaptability of our biophysical-based prediction approach.

7) Prediction Methodology

Moving ahead with creating a novel biophysical parameters-based exon-intron boundary prediction tool, we integrated the three benchmarked models into an easy-to-use programmed pipeline. This OS-independent pipeline, developed entirely in Python 3, is accessible as a command-line tool and is publicly available on GitHub. The exact methodology during a prediction involves validating the input sequence length and characters and then converting the input sequence into seven

numerical profiles corresponding to the combined major categories. Subsequently, a transient data frame with an organization similar to our training-testing dataset is created at the backend. This data frame then employs the organism-specific models and the reliability threshold value chosen by the user in addition to the sequence input step. The predictions from the employed models pass through various filters to provide the user with the final exon-intron boundary sites organized in an output file. The detailed prediction pipeline and all the filtering steps are available in the user manual **(Supplementary File 5)**. To test the working of the developed pipeline and its cross-platform functionality, predictions were made by ChemEXIN on varying-length genes from the three organisms in consideration.

Declarations

Data availability:

The datasets for all the studied organisms can be downloaded from the SCFBio website (http://www.scfbio-iitd.res.in/ChemEXIN/ChemEXIN_Datasets.tar).

Code availability:

ChemEXIN available as a Python-based command line utility is hosted at GitHub (https://github.com/rnsharma478/ChemEXIN). The archived source code is available at Zenodo, DOI: https://zenodo.org/doi/10.5281/zenodo.11101312.

Acknowledgements:

We express our sincere gratitude to Dr Modesto Orozco and Dr Federica Battistini from the Molecular Modelling and Bioinformatics group at the Institute of Research in Biomedicine, Barcelona, Spain, for sharing the structural MD simulation data utilized in this research. DS thanks the Council of Scientific & Industrial Research (CSIR), Government of India, for granting him the senior research fellowship. The authors acknowledge the Centre of Excellence Support to SCFBio, IIT Delhi from the Department of Biotechnology, Government of India.

Author Contributions:

The study's design was conceptualized by BJ and DS. Data collection, analysis, and development were conducted by DS, DA, and KS. DS, DA, and BJ analyzed the results and authored the manuscript. DS, DA, and KS developed and hosted the ChemEXIN on GitHub. AM made significant contributions to the conceptual framework of the investigation

Competing interests:

The authors declare no competing interests.

References

1. Spang, A., et al.: Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature. **521**(7551), 173–179 (2015)
2. Sharp, P.A.: Split genes and RNA splicing. Cell. **77**(6), 805–815 (1994)
3. Soller, M.: Pre-messenger RNA processing and its regulation: a genomic perspective. Cell. Mol. Life Sci. CMLS. **63**, 796–819 (2006)
4. Anna, A., Monika, G.: Splicing mutations in human genetic disorders: examples, detection, and confirmation. J. Appl. Genet. **59**, 253–268 (2018)
5. Mathé, C., Sagot, M.F., Schiex, T., Rouzé, P.: Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res. **30**(19), 4103–4117 (2002)

6. Allen, J.E., Pertea, M., Salzberg, S.L.: Computational gene prediction using multiple sources of evidence. Genome Res. **14**(1), 142–148 (2004)
7. Watson, J., Baker, T., Bell, S., et al.: *Molecular Biology of the Gene*. 7th edition. New York, Cold Spring Harbor Laboratory Press; ISBN-13: 978-0-321-76243-6 (2013)
8. Mishra, A., et al.: Intron exon boundary junctions in human genome have in-built unique structural and energetic signals. Nucleic Acids Res. **49**(5), 2674–2683 (2021)
9. Roca, X., Krainer, A.R.: Recognition of atypical 5' splice sites by shifted base-pairing to U1 snRNA. Nat. Struct. Mol. Biol. **16**(2), 176–182 (2009)
10. Roca, X., Sachidanandam, R., Krainer, A.R.: Intrinsic differences between authentic and cryptic 5' splice sites. Nucleic Acids Res. **31**(21), 6321–6333 (2003)
11. Nilsen, T.W., Graveley, B.R.: Expansion of the eukaryotic proteome by alternative splicing. Nature. **463**(7280), 457–463 (2010)
12. Senapathy, P., Shapiro, M.B., Harris, N.L.: [16] Splice junctions, branch point sites, and exons: Sequence statistics, identification, and applications to genome project (1990)
13. Brunak, S., Engelbrecht, J., Knudsen, S.: Prediction of human mRNA donor and acceptor sites from the DNA sequence. J. Mol. Biol. **220**(1), 49–65 (1991)
14. Yeo, G., Burge, C.B.: Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. In *Proceedings of the seventh annual international conference on Research in computational molecular biology* (pp. 322–331) (2003), April
15. Sahashi, K., et al.: In vitro and in silico analysis reveals an efficient algorithm to predict the splicing consequences of mutations at the 5' splice sites. Nucleic Acids Res. **35**(18), 5995–6003 (2007)
16. Ramakrishna, R., Srinivasan, R.: Gene identification in bacterial and organellar genomes using GeneScan. Comput. Chem. **23**(2), 165–174 (1999). http://hollywood.mit.edu/GENSCAN.html [Accessed 08-02-2024]
17. Yeh, R.F., Lim, L.P., Burge, C.B.: Computational inference of homologous gene structures in the human genome. Genome Res. **11**(5), 803–816 (2001)
18. Birney, E., Clamp, M., Durbin, R., GeneWise: genomewise Genome Res. **14**(5), 988–995 (2004)
19. Stanke, M., Morgenstern, B.: AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic acids research*, 33(suppl_2), W465-W467 URL: (2005). https://bioinf.uni-greifswald.de/augustus/submission.php [Accessed 14-02-2024]
20. Solovyev, V., Kosarev, P., Seledsov, I., Vorobyev, D.: Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome biology*, 7, 1–12 URL: (2006). http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind [Accessed 2024-02-08]
21. Snyder, E.E., Stormo, G.D.: Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. Nucleic Acids Res. **21**(3), 607–613 (1993)
22. Blanco, E., Parra, G., Guigó, R.: Using geneid to identify genes. Curr. protocols Bioinf. **18**(1), 4–3 (2007). https://github.com/guigolab/geneid https://github.com/guigolab/geneid/blob/master/README.md[Accessed 08-02-2024] [Accessed 08-02-2024]
23. Scalzitti, N., et al.: Spliceator: multi-species splice site prediction using convolutional neural networks. BMC Bioinform. **22**, 1–26 (2021). https://www.lbgi.fr/spliceator/ [Accessed 08-02-2024]
24. Trapnell, C., Pachter, L., Salzberg, S.L.: TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. **25**(9), 1105–1111 (2009)
25. Au, K.F., Jiang, H., Lin, L., Xing, Y., Wong, W.H.: Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Nucleic Acids Res. **38**(14), 4570–4578 (2010)

26. Wang, K., et al.: MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. **38**(18), e178–e178 (2010)
27. Ameur, A., Wetterbom, A., Feuk, L., Gyllensten, U.: Global and unbiased detection of splice junctions from RNA-seq data. Genome Biol. **11**, 1–9 (2010)
28. Levin, L., et al.: LEMONS–a tool for the identification of splice junctions in transcriptomes of organisms lacking reference genomes. PloS one. **10**(11), e0143329 (2015)
29. Fincher, J.A., Tyson, G.S., Dennis, J.H.: DNA-Encoded chromatin structural intron boundary signals identify conserved genes with common function. Int. J. Genomics, (2015)
30. Dickerson, R.E., Drew, H.R.: Structure of a B-DNA dodecamer: II. Influence of base sequence on helix structure. J. Mol. Biol. **149**(4), 761–786 (1981)
31. Yanagi, K., Privé, G.G., Dickerson, R.E.: Analysis of local helix geometry in three B-DNA decamers and eight dodecamers. J. Mol. Biol. **217**(1), 201–214 (1991)
32. El Hassan, M.A., Calladine, C.R.: The assessment of the geometry of dinucleotide steps in double-helical DNA; a new local calculation scheme. J. Mol. Biol. **251**(5), 648–664 (1995)
33. Olson, W.K., Gorin, A.A., Lu, X.J., Hock, L.M., Zhurkin, V.: B. DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proceedings of the National Academy of Sciences*, 95(19), 11163–11168 (1998)
34. Beveridge, D.L., et al.: Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d (CpG) steps. Biophys. J. **87**(6), 3799–3813 (2004)
35. Dixit, S.B., et al.: Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. Biophys. J. **89**(6), 3721–3740 (2005)
36. Lavery, R., Moakher, M.J.H.P.D., Maddocks, J.H., Petkeviciute, D., Zakrzewska, K.: Conformational analysis of nucleic acids revisited: Curves+. Nucleic Acids Res. **37**(17), 5917–5929 (2009)
37. Lavery, R., et al.: A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. Nucleic Acids Res. **38**(1), 299–313 (2010)
38. Pasi, M., et al.: μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. Nucleic Acids Res. **42**(19), 12272–12283 (2014)
39. Rohs, R., et al.: The role of DNA shape in protein–DNA recognition. Nature. **461**(7268), 1248–1253 (2009)
40. Florquin, K., Saeys, Y., Degroeve, S., Rouze, P., Van de Peer, Y.: Large-scale structural analysis of the core promoter in mammalian and plant genomes. Nucleic Acids Res. **33**(13), 4255–4264 (2005)
41. Gromiha, M.M., Siebers, J.G., Selvaraj, S., Kono, H., Sarai, A.: Intermolecular and intramolecular readout mechanisms in protein–DNA recognition. J. Mol. Biol. **337**(2), 285–294 (2004)
42. Mishra, A., Dhanda, S., Siwach, P., Aggarwal, S., Jayaram, B.: A novel method SEProm for prokaryotic promoter prediction based on DNA structure and energetics. Bioinformatics. **36**(8), 2375–2384 (2020)
43. Mishra, A., et al.: Toward a universal structural and energetic model for prokaryotic promoters. Biophys. J. **115**(7), 1180–1189 (2018)
44. Singhal, P., Jayaram, B., Dixit, S.B., Beveridge, D.L.: Prokaryotic gene finding based on physicochemical characteristics of codons calculated from molecular dynamics simulations. Biophys. J. **94**(11), 4173–4183 (2008)
45. Dutta, S., et al.: A physicochemical model for analyzing DNA sequences. J. Chem. Inf. Model. **46**(1), 78–85 (2006)
46. Khandelwal, G., Lee, R.A., Jayaram, B., Beveridge, D.L.: A statistical thermodynamic model for investigating the stability of DNA sequences from oligonucleotides to genomes. Biophys. J. **106**(11), 2465–2473 (2014)
47. Khandelwal, G., Bhyravabhotla, J.: A phenomenological model for predicting melting temperatures of DNA sequences. PloS one. **5**(8), e12433 (2010)

48. Sharma, D., et al.: Molecular dynamics simulation-based trinucleotide and tetranucleotide level structural and energy characterization of the functional units of genomic DNA. Phys. Chem. Chem. Phys. **25**(10), 7323–7337 (2023)
49. Singh, A., Mishra, A., Khosravi, A., Khandelwal, G., Jayaram, B.: Physico-chemical fingerprinting of RNA genes. Nucleic Acids Res. **45**(7), e47–e47 (2017)
50. Khandelwal, G., Jayaram, B.: DNA–water interactions distinguish messenger RNA genes from transfer RNA genes. J. Am. Chem. Soc. **134**(21), 8814–8816 (2012)
51. Nedelcheva-Veleva, M.N., et al.: The thermodynamic patterns of eukaryotic genes suggest a mechanism for intron–exon recognition. Nat. Commun. **4**(1), 2101 (2013)
52. Kraeva, R.I., et al.: Stability of mRNA/DNA and DNA/DNA duplexes affects mRNA transcription. PLoS One, **2**(3), e290 (2007)
53. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2012)
54. Frankish, A., et al.: GENCODE. *Nucleic acids research*, 49(D1), D916-D923 (2021). (2021)

Tables

Table 1. Comparison of methods using untrained non-protein coding *H. sapiens* gene (lncRNA, tRNA, and rRNA). sequences.

Method	Gene Category	TP[1]	FP[2]	TN[3]	FN[4]	Sensitivity	Specificity	F1-score	Precision	Accuracy
ChemEXIN	lncRNA	161	178	169	171	0.485	0.487	0.480	0.475	0.486
	tRNA	122	110	146	124	0.496	0.570	0.510	0.526	0.534
	rRNA	84	113	83	94	0.472	0.423	0.448	0.426	0.447
Augustus	lncRNA	110	330	268	504	0.179	0.448	0.209	0.250	0.312
	tRNA	82	88	292	328	0.200	0.768	0.283	0.482	0.473
	rRNA	88	90	184	414	0.175	0.672	0.259	0.494	0.351

[1] TP: True Positive.
[2] FP: False Positive.
[3] TN: True Negative.
[4] FN: False Negative.

Table 2. Performance evaluation of ChemEXIN on random genes from *H. sapiens* and *M. musculus*.

Organism	Gene	Length (nt[1])	Predicted Sites	Average time (sec)[2]
H. sapiens	*DMD*[3]	2,220,382	47	221.77
	BDNF[4]	188,307	28	24.84
	NEU1[5]	10,881	8	8.24
M. musculus	*RP1*[6]	409,685	26	41.30
	CDK6[7]	189,524	9	22.46
	SCAF8[8]	83,888	15	12.78

[1] nt: Nucleotides.
[2] Average time (sec): Average processing time calculated over three major OS (Windows version 10, Linux version Ubuntu 22.04, and macOS version 14 Sonoma) in Seconds.
[3] *DMD*: Dystrophin (muscular dystrophy, Duchenne and Becker types).

[4] *BDNF*: Brain-derived neurotrophic factor.
[5] *NEU1*: Neuraminidase-1.
[6] *RP1*: Retinitis Pigmentosa-1.
[7] *CDK6*: Cyclin-dependent kinase-6.
[8] *SCAF8*: SR-related CTD associated factor-8.

Figures

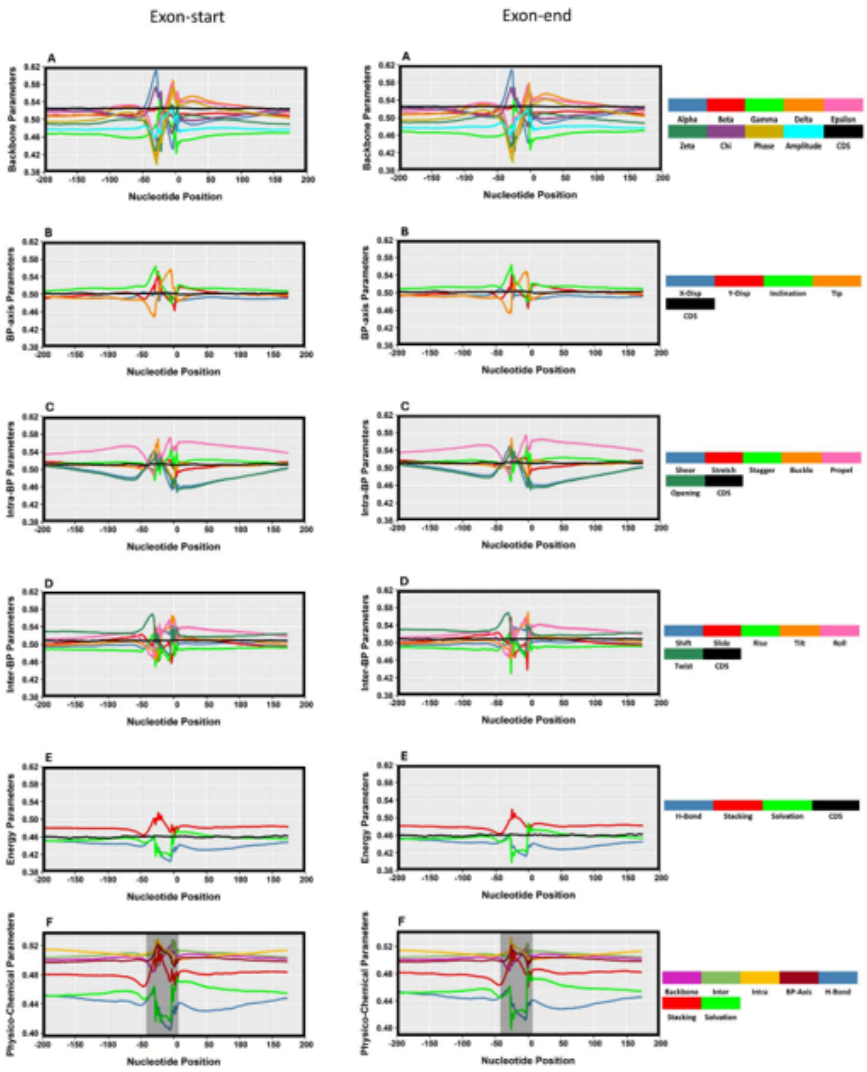

Figure 1

(A-E) Normalized structural and energy profiles at ES and EE. Individual parameter trends within each major category are shown separately, with normalized parameter values on the ordinate and nucleotide position relative to the ES/EE on the abscissa. (F) Combined structural and individual energy profiles plotted together, with the grey region highlighting the area undergoing transition.

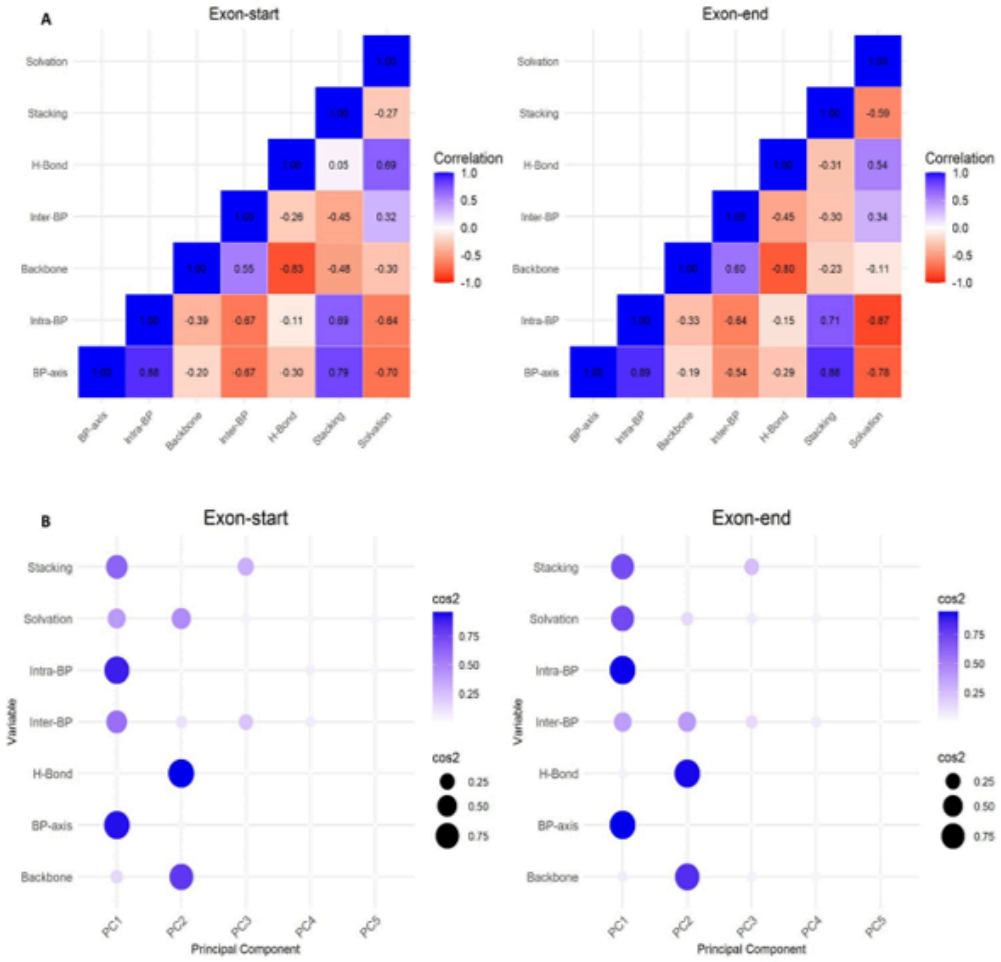

Figure 2

(A) Correlation matrices depicting the relationships among the seven final parameters at ES and EE. (B) Feature importance analysis conducted through PCA, revealing significant contributions of all seven parameters.

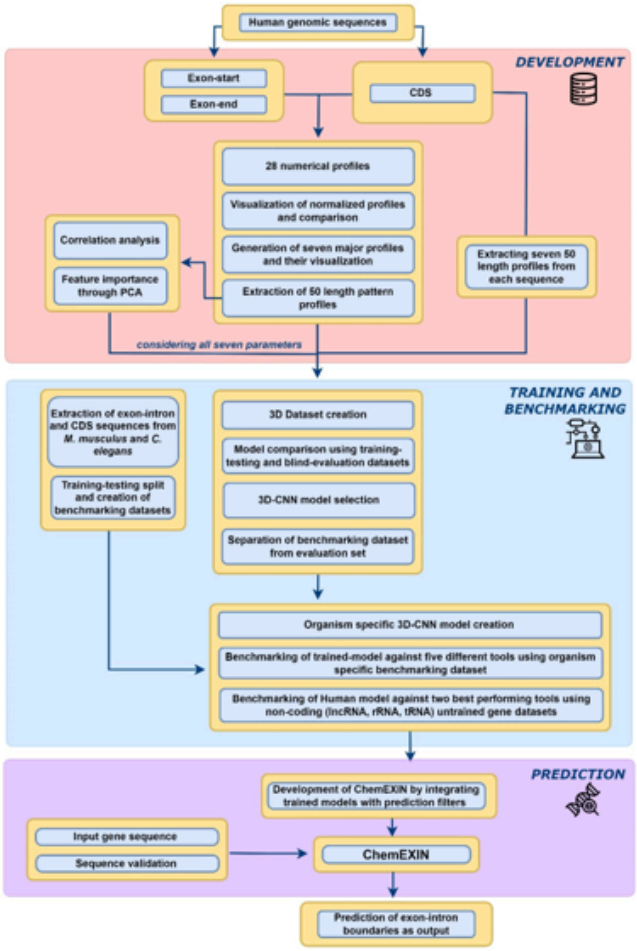

Figure 3

A flowchart outlining the steps implemented in ChemEXIN.

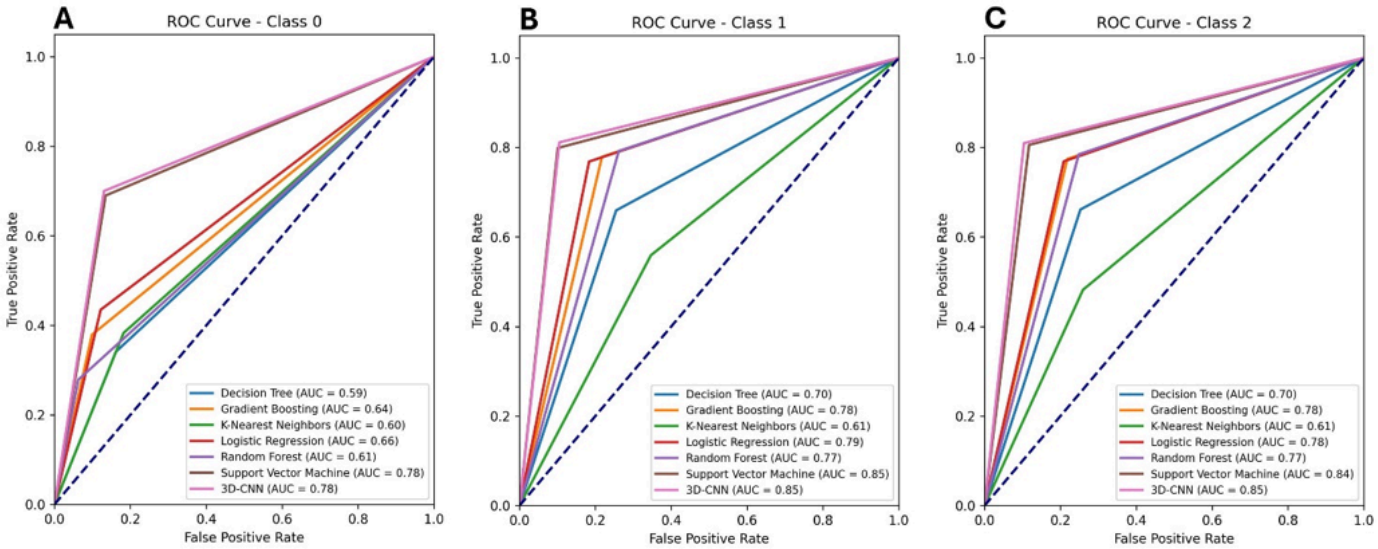

Figure 4

AUROC depicting AUC scores for all three classes (A) CDS:0, (B) ES:1, and (C) EE: 2 across all classifiers employed over the Blind-Evaluation Set

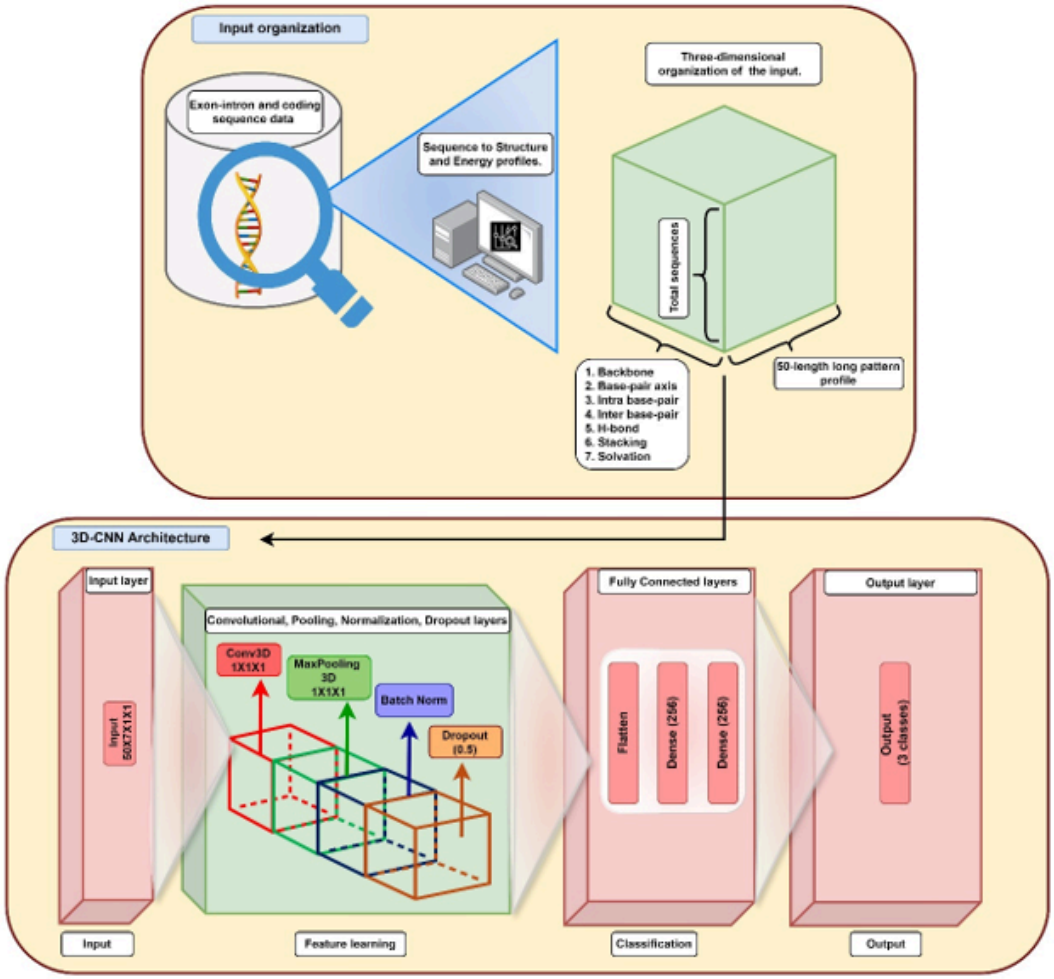

Figure 5

3D-CNN architecture employed within the three organism-specific models.

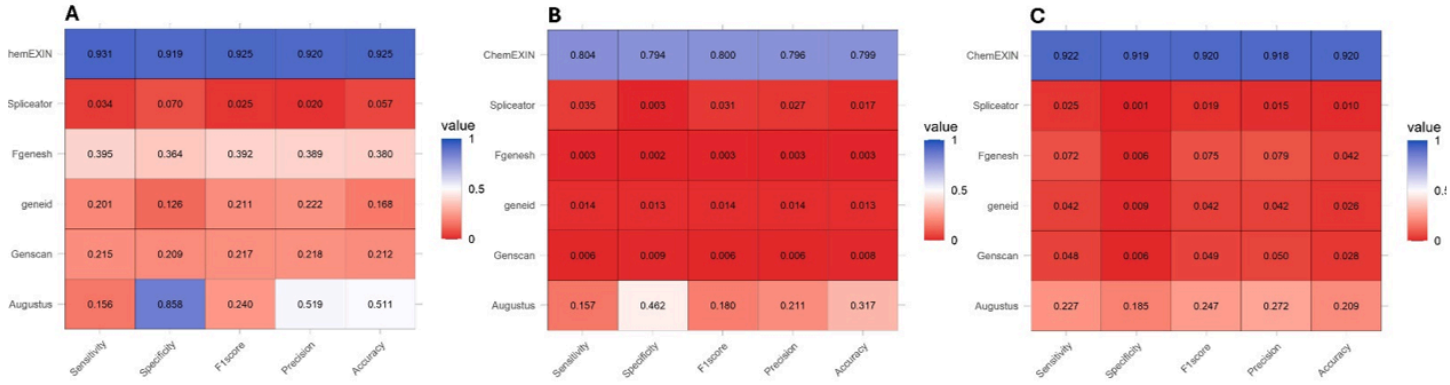

Figure 6

Heatmaps depicting the performance of all methods across all three organisms. (A) *H. sapiens* (B) *M. musculus* (C) *C. elegans*.

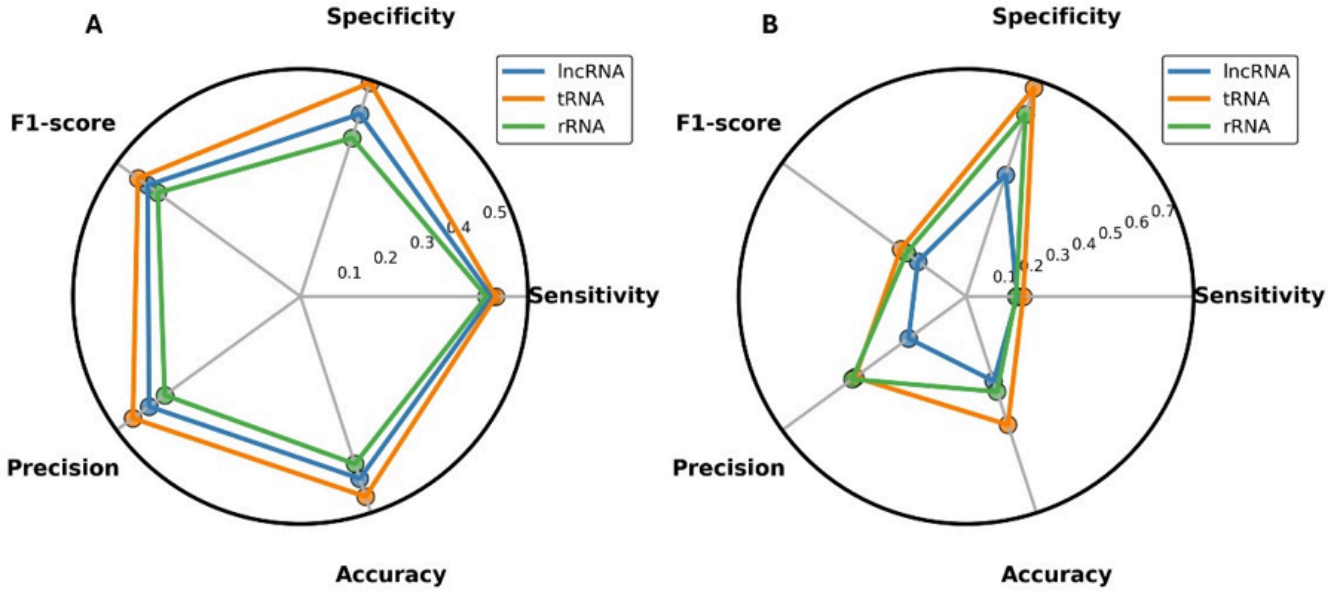

Figure 7

Performance evaluation of (A) ChemEXIN, and (B) Augustus on non-protein coding gene (lncRNA, tRNA, and rRNA) datasets.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile1.pdf
- SupplementaryFile2.pdf
- SupplementaryFile3.pdf
- SupplementaryFile4.pdf
- SupplementaryFile5.pdf